ISSN 2320-9984
Volume 3 Issue 2 | June 2015
ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

The Novel WFCM Algorithm for Dimensionality Reduction of High Dimensional Datasets

Rahul Patidar
Research Scholar,
Vaishnavi Institute of Technology and Science (VITS), Bhopal (M.P.), [INDIA]
Email: patidar_rahul@rediffmail.com

Sumit Sharma
Assistant Professor
Vaishnavi Institute of Technology and Science (VITS), Bhopal (M.P.), [INDIA]
Email: sumit_sharma782022@yahoo.co.in

Abstract*—*Dimensionality reduction studies techniques that successfully reduce data dimensionality for proficient data processing assignments such as pattern recognition (PR) machine learning (ML), text retrieval, and data mining (DM). From last many years Broad research into dimensionality reduction is being carried out and presently it's also in demand for additional growing due to imperative high-dimensional applications such as document indexing, text categorization, and gene expression data. In this article, we illustrated the novel concepts of dimensionality reduction, and in brief introduced some delegate techniques. We then presented in short some experiments of dimensionality reduction with six HD datasets from UCI repository. Our experimental study shows that proposed method is far better than compared previous methods.

Keywords:— *Dimensionality Reduction, Feature selection, Feature Extraction, FCM, WFCM, High Dimensional Dataset,*

1. INTRODUCTION

A dimension means to an extent of a definite features of an object. Dimensionality reduction (DR) is the research of techniques to reducing the number of dimensions to describing the object. Its common objectives are to eliminate unrelated and redundant data so that the computational cost will reduce and stay away from data over-fitting [1], and to extend the quality of data for well-organized data-intensive processing work such as pattern recognition (PR) and data mining (DM). Dimensionality reduction is a valuable solution to the dilemma of "curse of dimensionality". When the number of dimensions increases linearly, experiments have shown that the required number of examples for learning increases exponentially

Today practitioners and researchers interchangeably use attribute, dimension, feature, and variable. Likewise, we will interchangeably use example, object, instance, and vector. Consider an application in which a system processes data (images, speech signal or patterns in general) in the form of a collection of vectors. For a specific application, it is more regularly than not that a subset of features is significant and in some cases, a huge number of features are unrelated. This problem can be reasoned by factors such as:

- Many dimensions will have dissimilarity lesser than the measurement noise and thus will be irrelevant
- Many dimensions will be correlated (through functional dependence or linear combinations) to others and thus will be redundant.

Thus, in many circumstances, it is counseled to eliminate the irrelevant and

redundant dimensions, producing a more reasonable representation of the data [2].

Dimensionality reduction may be an analysis area at the intersection of many disciplines, as well as AI, statistics, databases, data mining (DM), pattern recognition, text mining, machine learning, visualization and optimization. Every of those areas have its own method of viewing the problem. As in case, in pattern recognition the trouble of dimensionality reduction is to remove a little set of features that improves most of the changeability of the info. In text mining, however, the problem is outlined as choosing a little set of words or terms (not new features that are combination of words or terms). Use of this vital technique conjointly varies with the applying domain. Samples of applications of dimensionality reduction techniques include: mining of text documents, gene structure discovery, image process, applied mathematical learning, and preliminary knowledge analysis. Data applications ought to be treated with different techniques. Betting on the applying, new features is also extracted as within the case of preliminary analysis, or a little set of original features is chosen as within the case of gene structure discovery.

Dimensionality reduction techniques are often sorted in numerous ways: (1) feature choice or feature extraction, (2) linear or non-linear, (3) supervised or unsupervised, and (4) local or global. Dimensionality reduction ways are usually classified into feature choice or feature extraction. In feature choice, a set of original options is chosen within the end. In feature extraction, new features are extracted using some mapping (linear or non-linear) from the first set of features. Linear techniques like principal components analysis (PCA) use a linear mapping to extract new options from original features [3]. Similarly, non-linear techniques like Sammon's mapping [4], locally linear embedding [5], and ISOMAP [6] use a non-linear mapping to extract new features. Supervised techniques will take advantage of any category info present within the data whereas unsupervised techniques don't use this category info. One limitation of the supervised techniques is that characteristic variables that describe samples of infrequent categories tend to be simply removed as results of dimensionality reduction creating use of the category distribution. Typically, supervised dimensionality reduction techniques are often more divided into local or global techniques. In a very native methodology, features area unit selected for every class of the category feature; just in case of a world methodology, features are selected for all classes. Among of these other ways of categorizing dimensionality reduction techniques, we are going to principally describe numerous techniques of dimensionality reduction in terms of feature extraction or feature selection.

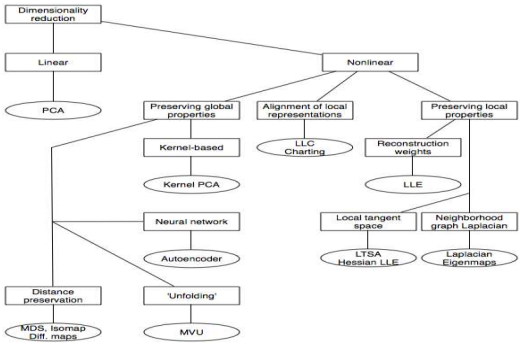

Figure 1: Classification of Dimension Reduction Technique

Figure 1 shows classification of techniques for dimensionality reduction. The main distinction between techniques for dimensionality reduction is the distinction between linear and nonlinear techniques. Linear techniques assume that the data lie on or near a linear subspace of the high-dimensional space. Nonlinear techniques for dimensionality reduction do not rely on the linearity assumption as a result of which more complex embeddings of the data in the high-dimensional space can be identified.

Organization the paper is, in section II we discussed few well recognized work done previously by researchers. In Section III we present our Proposed Novel WFCM algorithm with all major. In section IV experimental result analysis is shown with all comparative

studies with base work. Finally we concluded our paper in section V.

2. RELATED WORK

Our aim is efficient data clustering for high dimension datasets. Numerous researchers have developed many approaches for clustering high dimension datasets. Here, we explained some of the remarkable researches for clustering high dimension datasets.

Amorim [7] states a technique for clustering by using two pair-wise rules (must link and can't link) and a single-wise rules (cannot cluster) single-wise rule that uses very restricted amount of labeled information. they need demonstrated that the exactness of results might be improved by as well as these rules within the intelligent k-means algorithm and verified constant by means of experiments wherever the particular variety of clusters within the information has not been previously better-known to the method. Jun and Xiong [8], states a high-dimensional data clustering approach supported genetic algorithm, known as GA-HD clustering. Their clustering approach has known effective feature mathematical spaces by looking out the feature subspace using genetic algorithm. Binary encoded candidate features and cluster centres are used and also the extent of feature mathematical space contribution to mathematical space clustering has been proposed because the fitness functions. The utility and potency of the GAHD clustering algorithm are demonstrated by experimental results.

Khalilian et al. [9] have mentioned that dimension reduction by means of vertical data reduction performed before using clustering strategies for exceedingly large and high dimensional data sets has the most disadvantage of reducing the standard of results. Still, additional carefulness has been counseled because dimensionality reduction strategies unavoidably cause some loss of data or might impair the quality of the results, even disfiguring the important clusters. they need planned technique a technique to be used in high dimensional datasets that improves the performance of the K-Means clustering method by using divide and conquer technique with equivalency and compatible relation ideas.

The appropriate precision and speed up of their proposed technique have been verified by experimental results. Rajput *et al.* [10], suggested a basic framework by integrating the hypothesis of raw set theory (reduct) and k-means algorithm for proficient clustering of high dimensional data. Initially, by discarding the surplus attributes by means of the reduct idea of raw set theory, it has described the small dimensional space in the high dimensional (HD) data set. Then, it has described suitable clusters by grasping the k-means algorithm on this small dimensional data reduct. The actuality that the outline increases the effectiveness of the clustering process and the precision of the resulting clustering has been proved by their experiment on test dataset.

Tajunisha and Saravanan [11] stated that the preliminary centroids selected additionally because the dimension of the data considerably impact the worth of the ensuing clusters within the calculation costly k-means clustering algorithms used for a number of sensible applications. The exactness of the resultant value might have not been up to the mark once the dimensions of the dataset are high as a result of their undecided that the selected dataset is free from noise and defect. So, potency and exactness improvement necessitates decreasing the dimensionality of the given dataset. They need proposed a new technique that identifies initial centroids and conjointly, decreases the dimension of the data by using PCA to enhance the exactness of the cluster results.

Anaissi et al. [12] have planned a framework supported FS, linear dimensionality reduction and nonlinear dimensionality reduction for very high dimensional data reduction. they have proposed mutual info primarily based FS for screening features and

distinguishing the foremost applicable features with least redundancy. The potential variables have also, been extracted from a HD dataset by means of a kernel linear dimensionality reduction methodology. Also, the dimension has been reduced and therefore the data has been pictured employing a local linear embedding primarily based non-linear dimensionality reduction. Outputs of every step and therefore the efficiency of this framework are demonstrated by means of experimental results.

Dash et al. [13] have mentioned that preprocessing information by means that of an economical dimensionality reduction technique is crucial to enhance the efficiency and accuracy of the mining task on high dimensional data. They need simplified the analysis and image of multi dimensional knowledge set by employing a proposed PCA technique because the initial part for K-means clustering. They need additionally, created the formula more practical and efficient by distinguishing the initial centroids employing a new proposed technique. By examination the results of their proposed approach with that of the first approach, they need tested that their proposed approach obtains a lot of precise, straightforward to grasp results and most significantly takes significantly less time interval.

3. PROPOSED WORK

The practicality and weakness of using higher dimensional (HD) datasets in cluster algorithms are already explained above sections. The technique of dimension reduction was projected in [14] to overcome such sort of difficulties. but for streaming HD dataset then even once changing it into a less dimensional dataset the matter silently remains there [15], [14].

A. FCM algorithm

We take a dataset for example $X = \{x_1, x_2, x_3, ,,,,,, x_n\}$, the FCM algorithm divides X into c fuzzy clusters and determines every clusters center so that the objective function (cost function) of dissimilarity measure is reduced or less than a specific threshold. FCM inspect object value of each data in every cluster, this value is presented as follows:

Cost function:

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (d_{ik})^2$$

U and v can be estimated as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{(m-1)}}} \quad (2)$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \quad (3)$$

Wherei u_{ik} s the object value of thed k^{th} at x_k in thec i^{th} luster,| $d_{ik} =$ | x_k v_i | is the Euclidean distance among data x_k and the cluster centroid, v_i 1, i c 1, k n exponent> m 1. The FCM algorithm verifies the cluster centroids v_i and the object matrix U throughout the iterations using the following steps:

Initiate the object matrix, U u_{ik} at random arrives from (0, 1) and satisfy following:

$$\sum_{i=1}^{c}(u_{ik}) = 1, 1 \le k \le n$$

Calculate c fuzzy clusters= v_i i 1, ... c using Equation 3.

Calculate the objective function according to Equation 1. Terminate if Cost function of dissimilarity measure is lesser or deliberate on a particular value or if their

improvement result over previous iteration outcomes is less than a specific threshold or iterations arrive at a certain tolerance value.

Calculate a latest U using Equation 2. Go to step 2.

As FCM algorithm is clustering on the whole data set, and data stream might include a very huge data set, so to allow FCM algorithm to deal with data stream directly may employ significant amounts of CPU time to converge, or effect in an unbearable iteration numbers. For this condition, authors of [8] stated one alternative method called as weighted FCM algorithm (WFCM) for data stream as discussed as follow.

B. Weighted FCM (WFCM)

At first, separate data stream into chunks X_1, X_2,…… X_s with respect to the reaching time of data, and the range of each chunk is decided by primary storage space of the processing system, now let n_1, n_2,……n_s be the data numbers of chunks X_1, X_2,…… X_s correspondingly. Because of its stream setting, a time weight(w t) is imposed on each data representing the datum control amount on the clustering procedure, and

$$\int_{t_0}^{t_c} w(t)dt = 1$$

Where t_0 is initial time of stream and t_c is current time, the core idea of WFCM is renewing the concept of weighted clustering centers by iterations till the objective function gets a fulfilling result or the number of iteration is to a upper bound. Furthermore, during the procedure, as equation, we provide the singleton a constant weight. The procedure is represented as follow:

Bring in the chunk X_l (1) l s .
Revise the weight of cluster centroids.

If= l 1: employ FCM to gain cluster centroids= v_i i 1,…… c , and calculate:

$$w_i^{'} = \sum_{j=1}^{n_1}(u_{ij})w_j \qquad 1 \le i \le c$$

Where

$$w_j = 1, \forall 1 \le j \le n_1$$

Ifl l > 1:

$$w_i^{'} = \sum_{j=1}^{n_l+c}(u_{ij})w_j \qquad 1 \le i \le c$$

Where

$$w_j = 1, \forall c+1 \le j \le n_l + c$$

The centroids weight w_i then revises as

$$w_i = w_i^{'}$$

3) Revise cluster centroids:

$$v_i = \frac{\sum_{k=1}^{n_l+c} w_k (u_{ik})^m x_k}{\sum_{k=1}^{n_l+c} w_k (u_{ik})^m}$$

Where| $x_k \in v_i$ } $1 \le i \le c$

X_l **4)** Calculate objective function:

$$J_m(U,v) = \sum_{k=1} c + n_l \sum_{i=1}^{c} w_k (u_{ik})^m (d_{ik})^2$$

Stop if cost function is lesser or reached at a specific value, or its progress over preceding results obtained from iterations is less a specific threshold, or iterations reach a specific upper bound value.

5) Calculate a new U using Equation 2. Go to step 2.

6) If l s then stop, else go to step 1.

C. high dimensional dataset to two dimensional dataset conversions

The method suggested by authors [14] is used for reducing dimension of high dimensional (HD) datasets. In this technique

all high dimensional data presented in the dataset is altered into a two dimensional (2D) co-ordinate point. Therefore the imposed clustering algorithm is capable to obtain the converted two dimensional (2D) dataset as input as a substitute of higher dimensional (HD) dataset.

The operational steps of the dimension reduction method [14] are stated below:

Let $O = o_i, o_2, \ldots, o_n$ be a *d*-dimensional dataset. Now to alter each *d*-dimensional data $o_i \in O$ into two dimensional coordinate point (*Xi, Yi)* do the following:

Compute X_i and Y_i as

$$X_i = \frac{x_{i0} + x_{i1} + \ldots\ldots x_{id-1}}{d}$$

And

$$Y_i = \frac{y_{i0} + y_{i1} + \ldots\ldots y_{id-1}}{d}$$

For each j^{th} dimensional value of i^{th} data in *O* (i.e., o_{ij}), we could get a co-ordinate poin. (x_{ij}, y_{ij}) Whera $x_{ij} = r_{ij}\cos\theta_j$ nd. $y_{ij} = r_{ij}\sin\theta_j$ r_{ij} means the value of o_{ij} (value ind j^{th} imension of i^{th} data)., $\theta_j = \theta_{j-1} + 360/d$ an. $\theta_0 = 0^0$ Also for each data $o_i \in O$, $1 \le i \le n$ there must be d numbers of coordinate point (x_{ij}, y_{ij}), $1 \le i \le n$ a $1 \le j \le d$ nd with assist of these coordinate point (x_{ij}, y_{ij}) we can get the mean value. (X_i, Y_i) Now Plot all n numbers of the mean points on the two dimensional (2D) plane and after that employ clustering algorithm on the plotted mean points.

We give details the weakness of applying FCM algorithm on a dataset of streaming behavior. We merged both dimension reduction and WFCM method mutually to propose a novel clustering algorithm for high dimensional datasets. We call our recommend algorithm as Novel WFCM as we utilize the WFCM algorithm and dimension reduction (DR) technique for higher dimensional streaming datasets. Our algorithm is explained as follows:

Algorithm: Novel - WFCM

Input: High dimensional (*d*-dimensional) large dataset *O* of streaming behavior.

1) Translate the raw *d*-dimensional dataset *O* into two dimensional (2D) dataset *X* using the dimension reduction procedure as explained in section-C.

2) Employ the WFCM algorithm on the renewed two dimensional dataset *X* The WFCM algorithm is discussed in section B.

Remember it, because the raw dataset *O* has streaming in nature it is impossible to lessen the dimension of the entire dataset at a time. But it doesn't produce any trouble since WFCM algorithm utilizes a chunk of data from raw dataset at a time. We could clearly see it from section -B that before applying WFCM, we require to separate the raw dataset into number of data chunks. The purpose for this is for the reason that in real situation these data are streaming in nature and will not be uploaded into primary memory all together simultaneously. Therefore, the dimension reduction method has been employed on chunk basis and not all together

4. EXPERIMENTAL ANALYSIS

We consider higher dimensional dataset as input and transformed them into 2d (two dimensional) data set as mentioned in section III-C. Once reducing the dimension of the dataset we execute WFCM algorithm on that. Although WFCM already exists we utilize it here for clustering higher dimensional data to reducing their dimension. Experimental results

show that WFCM achieves higher than FCM for higher dimensional dataset of streaming behavior. Our main purpose is to specify that if we merge these two methods proposed in [14] and [15] along for a clustering algorithm then performance can acquire greater than before in contrast to the evaluation of any individual one. It is remarkable, our proposed algorithm (Novel-WFCM) is a hybrid procedure of the techniques proposed in [14] and [15]. For experimental study we utilize FCM algorithm on the reduced (2D) dataset as base algorithm and WFCM algorithm as the proposed. For the experiments we use six high dimensional datasets: connect-4, Forest fires, heart disease, iris, Soyabean, wine, these datasets are accessible in UCI repository

A. Cluster Validity

We suppose validity functions [8] to assessment of cluster efficiency. The cluster validity functions are supported on partition coefficient and partition entropy of *U*.

The Partition coefficient for FCM

$$V_{pc}(U) = \frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{2}\right)$$

The Partition coefficient for WFCM

$$V_{pc}(U) = \frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} w_i u_{ij}^{2}\right)$$

The Partition entropy for FCM

$$V_{pe}(U) = -\frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\log u_{ij}\right)$$

The Partition entropy for WFCM

$$V_{pe}(U) = -\frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} w_i u_{ij}\log u_{ij}\right)$$

Where i n s the total number of data in the dataset w_i , u_{ij} U are weight of centroids and membership matrix respectively. Table 1 and Table 2 represent results in detail with respect to six dataset we have taken for experiment.

Table 1: Cluster Validity (*Partition Coefficient) for all Six HD Datasets.*

Data Set	Connect-4		Forest Fires		Heart Dis-ease		Iris		Soyabean Large		Wine	
No of Cluster	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed
5	.53	.52	.73	.70	.89	.84	.86	.78	.75	.79	.74	.83
10	.61	.59	.60	.59	.90	.82	.75	.84	.68	.63	.72	.84
15	.62	.57	.52	.63	.98	.96	.70	.91	.71	.63	.67	.89
20	.59	.58	.51	.59	.97	.94	.64	.97	.63	.65	.67	.99
25	.58	.58	.48	.61	.98	.95	.61	1.1	.64	.67	.65	1.09

Table 2: Cluster Validity (*Partition entropy) for all six HD datasets.*

Data Set	Connect-4		Forest Fires		Heart Dis-ease		Iris		Soyabean Large		Wine	
No of Cluster	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed
5	95.3	62.5	53.21	39.67	18.96	14.64	29.60	39.29	39.46	26.39	50.82	30.12
10	94.5	63.5	86.75	57.38	19.77	18.19	53.56	51.82	62.66	59.01	61.72	46.53
15	103.04	70.09	110.48	67.63	39.75	33.4	66.17	64.63	52.50	70.74	74.89	56.80
20	112.12	70.73	117.89	84.79	45.77	38.7	78.96	75.28	75.75	76.77	78.55	59.15
25	117.23	72.88	129.60	88.74	54.6	45.3	90.04	70.84	69.92	86.74	84.47	59.94

Table 3: Memory Utilization (*in bytes) for all six HD datasets*

Data Set	Connect-4		Forest Fires		Heart Disease		Iris		Soyabean Large		Wine	
No of Cluster	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed
5	25941928	5188264	74632	14872	887704	177544	12200	2440	123696	24640	27096	5360
10	28644248	5728744	95392	19032	1627464	325504	18320	3680	138816	27680	34256	6800
15	31346568	6269224	116152	23192	2367224	473464	24440	4920	153936	30720	41416	8240
20	34048888	6809704	136912	27352	3106984	621424	30560	6160	169056	33760	48576	9680
25	**36751208**	7350184	157672	31512	3846744	769384	36680	7400	184176	36800	55736	11120

Table 4: Execution Time (*in seconds) for all six HD datasets*

Data Set	Connect-4		Forest Fires		Heart Disease		Iris		Soyabean Large		Wine	
No of Clus-ter	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed	Base	Pro-posed
5	9169	4941	700	137	1879	928	72	53	198	40	417	33
10	20485	1931	829	484	5868	1000	225	99	420	132	431	102
15	35823	5876	1523	818	18811	7411	301	128	392	185	1095	168
20	53943	9542	1969	1096	32717	12042	418	114	653	270	492	164
25	226141	13940	4097	1483	185513	17344	510	138	574	519	610	174

Memory Utilization:

Since WFCM method process as variety of chunks, we calculate the memory consumption of every chunk individually and consider the biggest value as the ultimate memory utilization for Novel - WFCM. Since the raw dataset is torrent or streaming in nature, it's not needed for Novel - WFCM to access more than one chunk simultaneously. Table 3 demonstrates the fraction of improvement in terms of memory utilization by proposed (Novel - WFCM) as compared to the FCM (base) algorithm. The improvement is 96% further better for all 6 datasets. FCM (Base algorithm) uses whole raw dataset at a time and therefore it needs enough memory to carry the whole dataset. This can be the rationale why base needs abundant higher memory than our proposed algorithm.

Execution Time

Similar as memory utilization we also computed execution time for each chunk separately and acquire the biggest value as the ultimate execution time for our proposed algorithm. Our ultimate aim is to compute the execution time of algorithm and Novel - WFCM would only consume one chunk at a instance and it is totally unknown that when the next chunk will appear. Table 4 shows the fraction of improvement in Novel - WFCM as compared to base FCM algorithm in terms of execution time. The very big improvement revealed because we compare the execution time of base algorithm (which utilizes whole raw dataset at a time) with the biggest execution time by a chunk in Novel - WFCM. The total execution time (with addition of the execution time of all the chunks) is also less than base algorithm but we have not shown it here.

5. CONCLUSION

As computers technologies become more and more powerful, several methods / applications will turn out large information of high dimensionality. Dimensionality reduction is an efficient method of dealing data with high dimensionality. The aim is to reduce the information so that process load decreases and patterns of higher quality may be extracted by pattern recognition and data mining algorithms. During this article, we delineated the novel ideas of dimensionality reduction, and shortly introduced some representative methods. We then presented in short some cases of dimensionality reduction for instance its application to several problem domains. Our experimental study shows that proposed methodology is far better than compared previous methods. The necessity of dimensionality reduction techniques presents new challenges, and novel methods are expected to be developed.

One future research direction is to increase these techniques to completely different application areas like microarray gene expression information. Generally micro array information has several genes however very less range of sample tests so suffering from the curse of dimensionality. Another analysis direction is to pick tuple and mix it with dimensionality reduction methodology. Typically researchers are performing dimensionality reduction and tuple selection severally. Another analysis direction includes Kernel PCA, probabilistic PCA, and independent component analysis.

REFERENCES:

[1] Ng. A. Y. Preventing overfitting of crossvalidation data. In Proceedings of Fourteenth International Conference on Machine Learning,pages 245–253, 1997

[2] U.M. Fayyad and R. Uthurusamy. Evolving data mining into solutions for insights. Communications of the Association for Computing Machinery, 45(8):28 – 31, August 2002.

[3] G. H. Dunteman. Principal Components Analysis. Sage Publications, 1989

[4] J. W. Sammon. A non-linear mapping for data structure analysis. IEEE Transactions on Computers, C-18 (5):401–409, 1969.

[5] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290:2323–2326, 2000.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2323, 2000

[7] de Amorim, Renato Cordeiro. "Constrained Intelligent K-Means: Improving Results with Limited Previous Knowledge." In Advanced Engineering Computing and Applications in Sciences, 2008. ADVCOMP'08. The Second International Conference on, pp. 176-180. IEEE, 2008.

[8] Sun, Hao-jun, and Lang-huan Xiong. "Genetic algorithm-based high-dimensional data clustering technique." In Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, vol. 1, pp. 485-489. IEEE, 2009.

[9] Khalilian M., Mustapha N., Suliman M., and Mamat D., “A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets,” in Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, vol. 1, pp. 503-507, 2010.

[10] Rajput, D., P. Singh, and Mahua

Bhattacharya. "An efficient and generic hybrid framework for high dimensional data clustering." In proceedings of International Conference on Data Mining and Knowledge Engineering (ICDMKE 2010), World Academy of Science, Engineering and Technology, Rome, pp. 174-179. 2010.

[11] Tajunisha, N., and V. Saravanan. "An increased performance of clustering high dimensional data using Principal Component Analysis." In Integrated Intelligent Computing (ICIIC), 2010 First International Conference on, pp. 17-21. IEEE, 2010.

[12] Anaissi, Ali, Paul J. Kennedy, and Madhu Goyal. "A framework for high dimensional data reduction in the microarray domain." In Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on, pp. 903-907. IEEE, 2010.

[13] Dash, B., Debahuti Mishra, A. Rath, and Milu Acharya. "A hybridized K-means clustering approach for high dimensional dataset." International Journal of Engineering, Science and Technology 2, no. 2 (2010): 59-66.

[14] P. Bishnu and V. Bhattacherjee, "A dimension reduction technique for k-means clustering algorithm," in Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, 2012, pp. 531–535

[15] R. Wan, X. Yan, and X. Su, "A weighted fuzzy clustering algorithm for data stream," in Proceedings of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management - Volume 01, ser. CCCM '08, 2008, pp. 360–364.